Міністерство освіти і науки України Національний технічний університет України «Київський політехнічний інститут» Фізико-технічний інститут

Лабораторна робота №1

З теми:

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Перевірила:

Селюх К. І.

Виконали:

студенти ІІІ курсу

групи ФБ-95

Гурджия Валерія

групи ФБ-94

Золотов Іван

Мета роботи: Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

Завдання:

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

2. За допомогою програми CoolPinkProgram оцінити значення (10) H , (20) H , (30) H .

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Код програми

Для написання даної лабораторної роботи ми використовували мову програмування python.

```
from array import *
from math import*


file_source = "..\\file1.txt"
file_noSpace = "..\\file2_noSpace.txt"
file_Space = "..\\file3_Space.txt"
file_probability_noSpace = "..\\file4_probability_noSpace.txt"
file_probability_Space = "..\\file5_probability_Space.txt"
file_bigramms_Space_1 = "..\\file6_bigramms_Space_1.txt"
file_bigramms_Space_2 = "..\\file7_bigramms_Space_2.txt"
file_bigramms_noSpace_1 = "..\\file8_bigramms_noSpace_1.txt"
file_bigramms_noSpace_2 = "..\\file9_bigramms_noSpace_2.txt"


Alphabet1=['а','б','в','г','д','е','ё','ж','з','и','й','к','л','м','н','о','п','р','с',
'т','у','ф','х','ц','ч','ш','щ','ы','ь','ъ','э','ю','я']
Alphabet2=['а','б','в','г','д','е','ё','ж','з','и','й','к','л','м','н','о','п','р','с',
'т','у','ф','х','ц','ч','ш','щ','ы','ь','ъ','э','ю','я', ' ']


file1 = open(file_source, "r")
cont_source = file1.read()

################################################################################
##############################   Открываем файл 2     ##########################

file2 = open(file_noSpace, "w")

# в этой функции мы меняем прописные буквы на строчные и убраем пробелы
def swap_case(s):
    swapped = []
    for char in s:
        if char.isalpha() == False:  # убраем все лишние символы
            char = ''

        if char.isupper():  # меняем прописные буквы на строчные
            swapped.append(char.lower())
        else:
            swapped.append(char)
    return ''.join(swapped)

file2.write(swap_case(cont_source))
file2.close()


###########  тут стираются буквы, которых нет в алфавите ##################
file2 = open(file_noSpace, "r")
cont_noSpace = file2.read()
file2.close()

file2 = open(file_noSpace, "w")
def alph(s):
    arr = []
    for char in cont_noSpace:
        if char == 'ё':
            char = 'е'
        if char == 'ъ':
            char = 'ь'

        if char not in Alphabet1:
            char = ''
```

```
        else:
            arr.append(char)
    return ''.join(arr)
file2.write(alph(cont_noSpace))
file2.close()


####################    Частота букв в файле 4     ##############################

file2 = open(file_noSpace, "r")
cont_noSpace = file2.read()

arr = []
thisset = set()
for char in cont_noSpace:
    thisset.add(char)
for c in thisset:
    arr.append(c)

arr.sort()
for i in range(0,len(arr)):
    k = (cont_noSpace.count(arr[i]))/len(cont_noSpace)
    strochka1 = arr[i] + "\t " +str(cont_noSpace.count(arr[i])) + "\t " +  str(k)
    file4 = open(file_probability_noSpace, "a")
    file4.write(strochka1 + '\n')
    file4.close()
file2.close()



################################################################################
############################   Открываем файл 3     ################################

file3 = open(file_Space, "w")

# в этой функции мы меняем прописные буквы на строчные
def swap_case(s):
    swapped = []
    for char in s:
        if char.isupper():  # меняем прописные буквы на строчные
            swapped.append(char.lower())
        else:
            swapped.append(char)
    return ''.join(swapped)


file3.write(swap_case(cont_source))
file3.close()

###########  тут стираются буквы, которых нет в алфавите ##################
file3 = open(file_Space, "r")
cont_Space = file3.read()
file3.close()

file3 = open(file_Space, "w")

def alph(s):
    arr = []
    for char in cont_Space:

        if char == '\n':
            char = ' '
        if char == 'ё':
            char = 'е'
        if char == 'ъ':
            char = 'ь'
```

```
        if char not in Alphabet2:
            char = ''
        else:
            arr.append(char)
    return ''.join(arr)

###########  тут стираются ненужные пробелы ##################
def space(s):
    s = s.split()
    return ' '.join(s)

string = alph(cont_Space)
file3.write(space(string))
file3.close()


####################   Частота букв в файле 3     ###############################

file3 = open(file_Space, "r")
cont_Space = file3.read()

arr = []
thisset = set()
for char in cont_Space:
    thisset.add(char)
for c in thisset:
    arr.append(c)


arr.sort()
for i in range(0,len(arr)):
    k = (cont_Space.count(arr[i]))/len(cont_Space)
    strochka2 = arr[i] + "\t " +str(cont_Space.count(arr[i])) + "\t " +  str(k)
    file5 = open(file_probability_Space, "a")
    file5.write(strochka2 + '\n')
    file5.close()
file3.close()

#########################################################################################
##############################     Биграммы    ########################################


############################  Файл 2 (без пробелов) с шагом 1      ###################

file2 = open(file_noSpace, "r")
cont_noSpace = file2.read()

arr1 = []
j = 0
for i in range(0,len(cont_noSpace)-1):
    char = cont_noSpace[i] + cont_noSpace[i+1]
    arr1.append(char)
    j+=1
arr = []
thisset = set()
for char in arr1:
    thisset.add(char)
for c in thisset:
    arr.append(c)
arr.sort()
for i in range(0, len(arr)):

    k = (cont_noSpace.count(arr[i]))/j
    strochka3 = arr[i] + "\t " +str(cont_noSpace.count(arr[i])) + "\t " +  str(k)
    file6 = open(file_bigramms_noSpace_1, "a")
    file6.write(strochka3 + '\n')
    file6.close()
```

```
file2.close()




##########################   Файл 2 (без пробелов) с шагом 2      ####################

file2 = open(file_noSpace, "r")
cont_noSpace = file2.read()

arr1 = []

j = 0
i = 0
while i < len(cont_noSpace)-1:
    char = cont_noSpace[i] + cont_noSpace[i+1]
    arr1.append(char)
    i+=2
    j+=1


arr = []
thisset = set()
for char in arr1:
    thisset.add(char)
for c in thisset:
    arr.append(c)

arr.sort()
for i in range(0, len(arr)):

    k = (cont_noSpace.count(arr[i]))/j
    strochka3 = arr[i] + "\t " +str(cont_noSpace.count(arr[i])) + "\t " +  str(k)
    file7 = open(file_bigramms_noSpace_2, "a")
    file7.write(strochka3 + '\n')
    file7.close()
file2.close()




##########################   Файл 3 (с пробелами) с шагом 1      ####################

file3 = open(file_Space, "r")
cont_Space = file3.read()

arr1 = []
j = 0
for i in range(0,len(cont_Space)-1):
    char = cont_Space[i] + cont_Space[i+1]
    arr1.append(char)
    j+=1
arr = []
thisset = set()
for char in arr1:
    thisset.add(char)
for c in thisset:
    arr.append(c)

arr.sort()

for i in range(0, len(arr)):

    k = (cont_Space.count(arr[i]))/j
    strochka3 = arr[i] + "\t " +str(cont_Space.count(arr[i])) + "\t " +  str(k)
```

```
    file8 = open(file_bigramms_Space_1, "a")
    file8.write(strochka3 + '\n')
    file8.close()
file3.close()




#######################   Файл 3 (с пробелами) с шагом 2      ######################

file3 = open(file_Space, "r")
cont_Space = file3.read()

arr1 = []

j = 0
i = 0
while i < len(cont_Space)-1:
    char = cont_Space[i] + cont_Space[i+1]
    arr1.append(char)
    i+=2
    j += 1

arr = []
thisset = set()
for char in arr1:
    thisset.add(char)
for c in thisset:
    arr.append(c)

arr.sort()
for i in range(0, len(arr)):

    k = (cont_Space.count(arr[i]))/j
    strochka3 = arr[i] + "\t " +str(cont_Space.count(arr[i])) + "\t " +  str(k)
    file9 = open(file_bigramms_Space_2, "a")
    file9.write(strochka3 + '\n')
    file9.close()

file3.close()
```

Скріншоти виконання

Ймовірність та кількість букв у файлі без пробілів

1		366268	0.16207
2	а	151089	0.06686
3	б	34940	0.01546
4	в	77582	0.03433
5	г	35768	0.01583
6	д	59477	0.02632
7	е	149450	0.06613
8	ж	18084	0.008
9	з	31574	0.01397
10	и	133692	0.05916
11	й	20844	0.00922
12	к	68503	0.03031
13	л	107360	0.04751
14	м	58658	0.02596
15	н	119586	0.05292
16	о	213328	0.0944
17	п	53240	0.02356
18	р	86558	0.0383
19	с	96233	0.04258
20	т	115078	0.05092
21	у	57106	0.02527
22	ф	2841	0.00126
23	х	15821	0.007
24	ц	6572	0.00291
25	ч	29917	0.01324
26	ш	15765	0.00698
27	щ	6245	0.00276
28	ы	33682	0.0149
29	ь	37195	0.01646
30	э	9714	0.0043
31	ю	10692	0.00473
32	я	37024	0.01638
33			

Ймовірність та кількість букв у файлі з пробілами

1		366268	0.16207
2	а	151089	0.06686
3	б	34940	0.01546
4	в	77582	0.03433
5	г	35768	0.01583
6	д	59477	0.02632
7	е	149450	0.06613
8	ж	18084	0.008
9	з	31574	0.01397
10	и	133692	0.05916
11	й	20844	0.00922
12	к	68503	0.03031
13	л	107360	0.04751
14	м	58658	0.02596
15	н	119586	0.05292
16	о	213328	0.0944
17	п	53240	0.02356
18	р	86558	0.0383
19	с	96233	0.04258
20	т	115078	0.05092
21	у	57106	0.02527
22	ф	2841	0.00126
23	х	15821	0.007
24	ц	6572	0.00291
25	ч	29917	0.01324
26	ш	15765	0.00698
27	щ	6245	0.00276
28	ы	33682	0.0149
29	ь	37195	0.01646
30	э	9714	0.0043
31	ю	10692	0.00473
32	я	37024	0.01638
33			

Частота біграм з шагом 1 у файлі без пробілів

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ы	ь	э	ю	я
а	771	3799	8776	2574	5530	4011	2559	7545	2531	2382	12804	22621	7196	10509	3000	6808	6450	11819	11105	1622	522	2592	475	3068	1463	788	0	0	859	2428	4381
б	2151	143	102	26	45	7008	41	17	5039	0	381	1977	55	739	5452	30	2770	339	39	2032	0	117	7	15	27	168	4674	298	235	27	979
в	11959	571	838	917	1351	10565	147	1284	6985	0	1766	1402	714	3998	12124	1898	1821	6303	1792	1771	75	141	88	388	1333	49	5168	999	502	17	600
г	3116	156	170	208	3093	1455	10	69	1596	2	327	3759	52	1035	16237	224	2287	244	131	1372	17	13	4	103	17	1	2	0	43	0	15
д	9421	150	2543	170	1847	11128	1308	130	6413	3	746	1308	254	3994	7453	525	2616	933	449	3938	10	128	690	80	125	2	1511	678	107	124	689
е	537	4408	7314	7688	6755	4012	1988	4352	2017	4057	4539	12299	9143	17940	3236	5708	16971	11118	13389	1390	270	1487	688	3021	2129	1228	0	0	580	198	795
ж	3265	52	57	62	1568	6121	33	26	3064	0	303	20	40	1849	599	64	33	59	42	520	5	10	0	179	1	0	0	84	17	0	11
з	11136	392	2022	828	1889	967	278	301	1215	0	794	617	961	3406	1655	473	877	404	285	1072	16	29	7	104	72	7	894	390	103	3	363
и	746	3405	8788	2526	5304	5094	869	6067	3572	2170	7606	16878	6721	11537	4219	5900	3335	10366	9369	1527	368	3776	2064	5241	1808	276	0	0	881	652	2447
й	223	896	1238	562	1124	251	203	495	979	23	2452	306	947	1440	1033	1634	844	2637	1213	466	152	140	205	713	281	27	0	0	187	17	156
к	15746	541	1144	312	534	2190	197	244	7765	0	821	2389	415	2253	18402	1011	5276	1429	2094	4542	45	148	109	291	91	15	1	0	330	10	148
л	14900	1438	1673	928	815	10332	667	570	18803	2	2164	3624	887	2729	15082	1666	974	6085	982	3547	114	172	22	1545	104	25	2282	8953	535	1857	3870
м	8290	857	1372	796	761	7398	145	406	7174	1	1379	749	826	4090	9306	2029	650	1874	741	4827	88	154	76	788	106	9	2065	230	383	48	1024
н	22532	561	931	616	1152	18417	118	641	15832	0	1294	341	381	5939	21667	1552	1414	2087	1799	7539	69	215	650	705	73	160	6604	2566	134	433	3152
о	319	8975	17287	11336	13133	4667	4770	4485	3334	8146	7107	14196	15118	20704	4493	7127	13695	18237	20095	1851	348	1641	303	5259	2341	637	0	0	994	272	2320
п	4396	125	16	14	8	4366	2	6	2199	0	320	2024	19	551	22418	197	12077	174	467	1991	2	4	5	56	33	0	826	115	21	18	774
р	15074	218	1273	354	1264	12356	747	140	13661	1	1224	1355	767	2733	15610	412	877	919	1838	6437	41	454	171	242	418	70	4137	1297	330	246	1886
с	2901	505	2456	314	989	6848	203	184	4267	0	6909	7082	2938	2836	5734	4322	673	2292	22481	1547	83	477	96	624	180	6	905	7844	241	318	9920
т	12050	690	5649	311	953	10735	252	277	8709	0	2460	693	732	3067	32771	1298	7342	3736	917	3160	79	126	243	771	86	15	4019	11501	913	120	1393
у	391	2736	3277	2522	3849	574	2617	1052	1389	285	3270	5348	3684	2530	960	2730	1485	4223	4142	476	111	1123	30	3028	1176	700	0	0	535	2504	306
ф	328	17	4	7	5	283	0	4	537	0	16	96	11	11	750	7	277	38	29	281	45	0	1	22	2	0	29	17	16	5	2
х	1271	269	1193	216	413	269	68	166	831	0	509	565	401	1064	4885	752	570	698	337	418	65	66	23	128	141	6	0	16	420	2	56
ц	1562	35	203	29	19	1764	5	29	977	0	45	7	27	69	638	72	25	54	72	389	0	6	2	9	2	0	511	0	13	0	8
ч	4340	19	29	5	24	6486	1	13	5506	0	1022	53	14	2003	198	26	14	22	7632	1652	7	2	1	84	266	0	4	476	10	0	3
ш	2024	7	127	0	7	3800	0	4	2847	0	1430	1448	95	501	763	50	61	11	300	502	1	1	21	4	27	0	0	1688	15	12	2
щ	880	0	1	1	0	2716	0	0	1966	0	0	0	0	90	8	3	3	1	0	533	0	0	0	0	0	0	0	43	0	0	0
ы	102	1037	3117	892	841	2919	166	569	690	3420	825	2491	3119	1259	739	1641	1015	2299	1784	356	20	1914	26	454	1541	17	40	0	219	4	156
ь	234	1089	2568	547	1184	1697	141	990	1746	6	3456	505	1239	5047	1899	2173	538	3895	1278	599	139	162	413	1456	1599	22	0	0	495	1051	1027
э	1	18	12	52	1686	0	1	7	0	134	211	197	91	1316	3	24	371	72	5297	0	21	34	0	0	13	0	0	0	117	0	1
ю	99	774	416	196	915	89	116	183	375	102	497	208	308	474	411	569	310	810	1093	107	53	125	32	588	107	1403	0	0	90	124	114
я	223	1050	2970	750	2415	739	432	1304	1493	110	1816	2789	1487	3861	1445	2299	901	2996	3876	589	74	561	120	946	186	614	0	0	354	198	421

Частота біграм з шагом 1 у файлі з пробілами

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ы	ь	э	ю	я
	0	3716	17252	31125	10018	16840	8748	2987	10605	21133	38	21975	5670	14850	32798	27039	37753	11715	31905	17504	10293	1587	3566	777	12039	2503	337	0	0	7389	119	3987
а	39434	336	1748	5724	1547	3926	2906	2345	6450	290	2375	10232	22060	5451	6600	46	2499	5288	8564	9194	646	323	2259	396	2028	1177	746	0	0	28	2407	3966
б	623	2136	100	74	14	29	6992	40	5	5001	0	368	1974	39	706	5383	0	2764	304	28	2003	0	114	6	8	26	168	4674	298	59	27	971
в	13645	11776	17	204	43	467	10306	10	1028	6519	0	502	1112	254	3101	11359	435	1353	5130	746	1525	1	55	40	90	1160	13	5168	999	2	9	505
г	1726	3103	9	28	182	2984	1410	0	6	1511	0	251	3746	12	870	16099	0	2238	44	76	1327	10	5	0	81	14	0	2	0	24	0	0
д	2744	9392	65	2357	35	1724	11046	1290	35	6284	1	513	1257	137	3667	7323	296	2516	652	313	3876	2	104	679	26	110	0	1511	678	51	121	668
е	35839	219	2686	4292	6743	5080	3480	1725	2995	144	4057	2824	11709	7398	15125	299	1871	15819	7650	11758	252	156	1025	623	1843	1852	1206	0	0	3	188	506
ж	629	3262	22	9	40	1534	6110	29	2	3017	0	278	18	10	1779	547	0	16	14	7	504	0	0	0	172	0	0	0	84	1	0	0
з	3980	11097	260	1748	656	1663	888	262	186	1056	0	452	554	818	3022	1409	98	731	23	52	963	0	0	1	10	14	0	894	390	5	2	326
и	45738	309	1050	4385	1475	3326	3927	615	4785	1156	2162	4943	16249	5119	7397	634	768	1877	6749	7120	36	136	3336	1998	4125	1491	239	0	0	8	644	1841
й	15596	13	79	98	12	349	49	1	49	12	21	1322	47	194	487	135	44	73	1071	652	3	33	3	113	306	72	2	0	0	2	2	4
к	10617	15651	2	392	8	21	2011	26	17	7002	0	277	2223	8	1272	17617	4	4912	482	1543	4207	6	7	76	6	30	0	1	0	69	4	3
л	20263	14712	137	29	221	106	9444	590	128	17711	0	935	3461	92	835	13571	38	1	4564	290	3187	45	6	1	213	14	11	2282	8953	26	1854	3639
м	16165	8062	137	35	270	14	7141	0	0	6110	0	252	498	195	2886	8083	327	152	288	178	4338	20	9	21	141	1	2	2065	230	150	37	837
н	9876	22464	189	25	450	769	18262	76	238	15125	0	964	258	8	5021	21028	14	1166	1121	1452	6966	52	47	644	489	19	157	6604	2566	12	433	3082
о	44492	45	6918	13367	10323	11180	3533	4239	3053	1282	8143	4839	13398	13261	16712	884	2754	12375	14108	17571	556	211	1169	196	3806	2124	571	0	0	109	267	1721
п	231	4386	113	1	7	3	4365	0	0	2179	0	306	2024	1	538	22407	179	12053	159	456	1980	1	2	5	49	32	0	826	115	17	18	772
р	2455	15040	102	1056	268	1141	12316	742	84	13436	0	997	1328	678	2511	15459	150	824	744	1759	6382	24	386	167	202	399	68	4137	1297	293	246	1862
с	6588	2810	163	2048	91	598	6702	156	21	4022	0	6402	6942	2610	2141	5304	3761	452	1878	21965	1351	50	415	86	509	132	5	905	7844	64	312	9871
т	12196	11910	178	4599	26	255	10349	0	9	8012	0	1756	480	198	1988	31824	242	6888	2728	388	2890	24	36	226	313	38	10	4019	11501	617	118	1252
у	13805	230	2190	2177	2243	3294	354	2543	754	277	284	2429	5181	3083	994	61	1385	1138	3138	3518	185	53	994	5	2028	1107	687	0	0	269	2503	150
ф	103	325	5	0	0	0	280	0	0	533	0	10	94	8	5	743	0	265	29	27	279	44	0	0	22	1	0	29	17	14	5	2
х	6757	1183	3	653	11	3	169	0	0	413	0	13	390	92	599	4390	1	326	86	64	255	0	13	1	0	43	0	0	16	337	0	0
ц	668	1551	0	146	0	0	1747	0	0	923	0	13	1	0	3	590	2	0	0	46	368	0	0	2	0	0	0	511	0	1	0	0
ч	258	4339	0	8	0	0	6483	0	0	5489	0	1012	49	7	1985	184	0	10	1	7615	1644	1	1	0	77	265	0	4	476	4	0	0
ш	196	2021	2	108	0	0	3798	0	0	2832	0	1408	1447	88	476	746	30	53	2	294	493	0	0	21	1	25	0	0	1688	7	12	0
щ	13	880	0	0	0	0	2716	0	0	1964	0	0	0	0	88	5	0	3	0	0	533	0	0	0	0	0	0	43	0	0	0	0
ы	10103	0	670	2105	697	308	2670	112	209	33	3420	365	2394	2737	261	2	419	774	1348	1322	8	0	1801	16	278	1503	14	40	0	0	0	63
ь	23603	1	175	151	74	202	929	2	304	102	0	2127	0	366	2818	35	5	0	1885	198	0	38	0	369	497	1484	8	0	0	0	1047	775
э	11	1	17	11	52	1686	0	1	7	0	134	210	197	91	1314	1	22	371	72	5296	0	21	34	0	0	13	0	0	0	116	0	1
ю	5552	16	545	19	27	636	4	46	64	22	101	102	106	85	127	0	21	99	368	776	0	2	97	16	300	39	1397	0	0	1	120	0
я	22362	5	100	600	226	1335	232	247	526	48	108	417	2492	754	1451	0	107	301	1090	2862	9	0	334	87	253	60	604	0	0	1	193	217

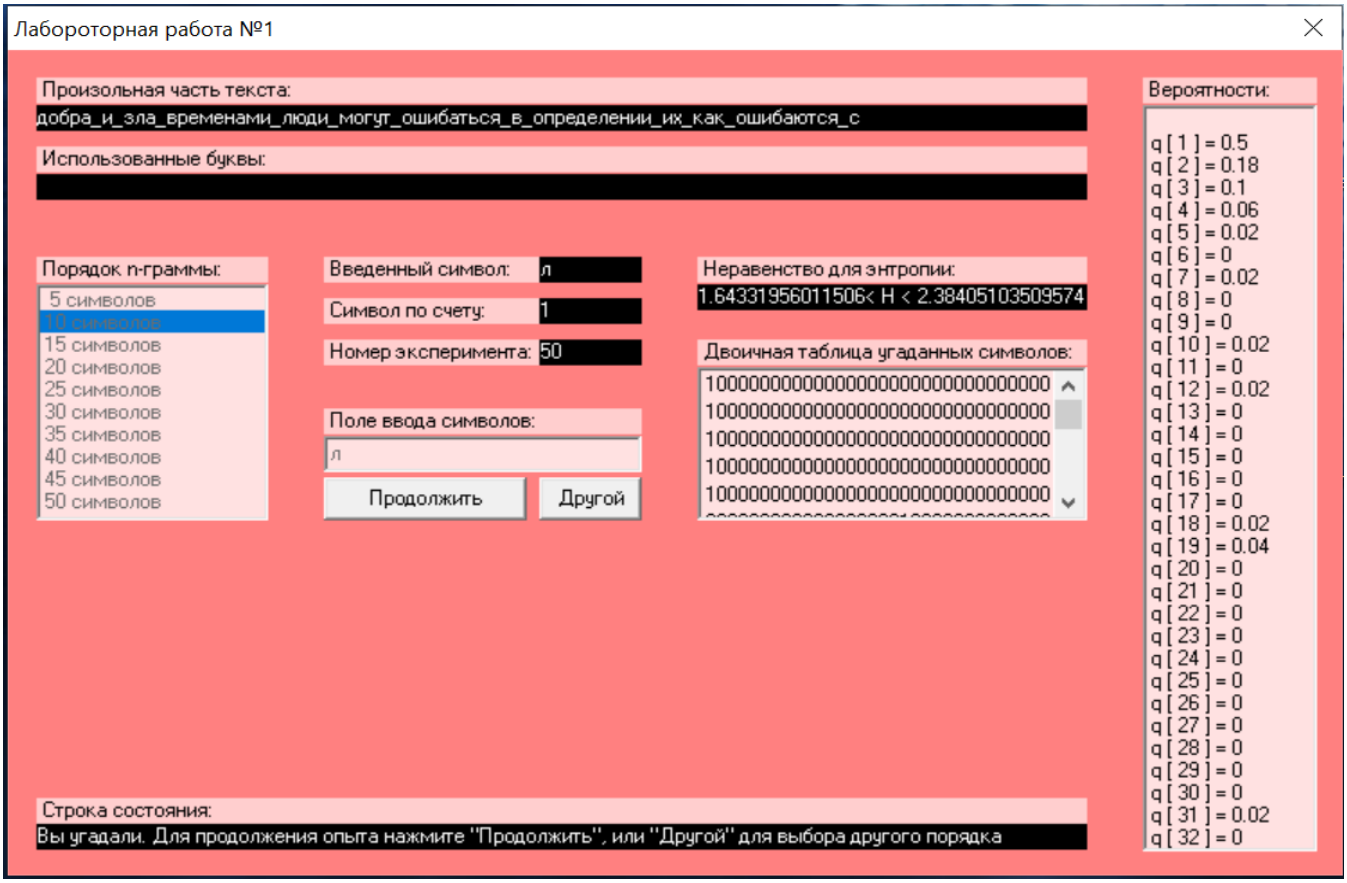
Частота біграм з шагом 2 у файлі без пробілів

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ы	ь	э	ю	я
а	771	3799	8776	2574	5530	4011	2559	7545	2531	2382	12804	22621	7196	10509	3000	6808	6450	11819	11105	1622	522	2592	475	3068	1463	788	0	0	859	2428	4381
б	2151	143	102	26	45	7008	41	17	5039	0	381	1977	55	739	5452	30	2770	339	39	2032	0	117	7	15	27	168	4674	298	235	27	979
в	11959	571	838	917	1351	10565	147	1284	6985	0	1766	1402	714	3998	12124	1898	1821	6303	1792	1771	75	141	88	388	1333	49	5168	999	502	17	600
г	3116	156	170	208	3093	1455	10	69	1596	2	327	3759	52	1035	16237	224	2287	244	131	1372	17	13	4	103	17	0	2	0	43	0	15
д	9421	150	2543	170	1847	11128	1308	130	6413	3	746	1308	254	3994	7453	525	2616	933	449	3938	10	128	690	80	125	2	1511	678	107	124	689
е	537	4408	7314	7688	6755	4012	1988	4352	2017	4057	4539	12299	9143	17940	3236	5708	16971	11118	13389	1390	270	1487	688	3021	2129	1228	0	0	580	198	795
ж	3265	52	57	62	1568	6121	33	26	3064	0	303	20	40	1849	599	64	33	59	42	520	5	10	0	179	1	0	0	84	17	0	11
з	11136	392	2022	828	1889	967	278	301	1215	0	794	617	961	3406	1655	473	877	404	285	1072	16	29	7	104	72	7	894	390	103	3	363
и	746	3405	8788	2526	5304	5094	869	6067	3572	2170	7606	16878	6721	11537	4219	5900	3335	10366	9369	1527	368	3776	2064	5241	1808	276	0	0	881	652	2447
й	223	896	1238	562	1124	251	203	495	979	23	2452	306	947	1440	1033	1634	844	2637	1213	466	152	140	205	713	281	27	0	0	187	17	156
к	15746	541	1144	312	534	2190	197	244	7765	0	821	2389	415	2253	18402	1011	5276	1429	2094	4542	45	148	109	291	91	15	0	0	330	10	148
л	14900	1438	1673	928	815	10332	667	570	18803	2	2164	3624	887	2729	15082	1666	974	6085	982	3547	114	172	22	1545	104	25	2282	8953	535	1857	3870
м	8290	857	1372	796	761	7398	145	406	7174	1	1379	749	826	4090	9306	2029	650	1874	741	4827	88	154	76	788	106	9	2065	230	383	48	1024
н	22532	561	931	616	1152	18417	118	641	15832	0	1294	341	381	5939	21667	1552	1414	2087	1799	7539	69	215	650	705	73	160	6604	2566	134	433	3152
о	319	8975	17287	11336	13133	4667	4770	4485	3334	8146	7107	14196	15118	20704	4493	7127	13695	18237	20095	1851	348	1641	303	5259	2341	637	0	0	994	272	2320
п	4396	125	16	14	8	4366	2	6	2199	0	320	2024	19	551	22418	197	12077	174	467	1991	2	4	5	56	33	0	826	115	21	18	774
р	15074	218	1273	354	1264	12356	747	140	13661	1	1224	1355	767	2733	15610	412	877	919	1838	6437	41	454	171	242	418	70	4137	1297	330	246	1886
с	2901	505	2456	314	989	6848	203	184	4267	0	6909	7082	2938	2836	5734	4322	673	2292	22481	1547	83	477	96	624	180	6	905	7844	241	318	9920
т	12050	690	5649	311	953	10735	252	277	8709	0	2460	693	732	3067	32771	1298	7342	3736	917	3160	79	126	243	771	86	15	4019	11501	913	120	1393
у	391	2736	3277	2522	3849	574	2617	1052	1389	285	3270	5348	3684	2530	960	2730	1485	4223	4142	476	111	1123	30	3028	1176	700	0	0	535	2504	306
ф	328	17	4	7	5	283	0	4	537	0	16	96	11	11	750	7	277	38	29	281	45	0	0	22	2	0	29	17	16	5	2
х	1271	269	1193	216	413	269	68	166	831	0	509	565	401	1064	4885	752	570	698	337	418	65	66	23	128	141	6	0	16	420	2	56
ц	1562	35	203	29	19	1764	5	29	977	0	45	7	27	69	638	72	25	54	72	389	0	6	2	9	2	0	511	0	13	0	8
ч	4340	19	29	5	24	6486	0	13	5506	0	1022	53	14	2003	198	26	14	22	7632	1652	7	2	0	84	266	0	4	476	10	0	3
ш	2024	7	127	0	7	3800	0	4	2847	0	1430	1448	95	501	763	50	61	11	300	502	0	1	21	4	27	0	0	1688	15	12	0
щ	880	0	0	0	0	2716	0	0	1966	0	0	0	0	90	8	3	3	0	0	533	0	0	0	0	0	0	0	43	0	0	0
ы	102	1037	3117	892	841	2919	166	569	690	3420	825	2491	3119	1259	739	1641	1015	2299	1784	356	20	1914	26	454	1541	17	40	0	219	4	156
ь	234	1089	2568	547	1184	1697	141	990	1746	6	3456	505	1239	5047	1899	2173	538	3895	1278	599	139	162	413	1456	1599	22	0	0	495	1051	1027
э	0	18	12	52	1686	0	0	7	0	134	211	197	91	1316	3	24	371	72	5297	0	21	34	0	0	13	0	0	0	117	0	1
ю	99	774	416	196	915	89	116	183	375	102	497	208	308	474	411	569	310	810	1093	107	53	125	32	588	107	1403	0	0	90	124	114
я	223	1050	2970	750	2415	739	432	1304	1493	110	1816	2789	1487	3861	1445	2299	901	2996	3876	589	74	561	120	946	186	614	0	0	354	198	421

Частота біграм з шагом 2 у файлі з пробілами

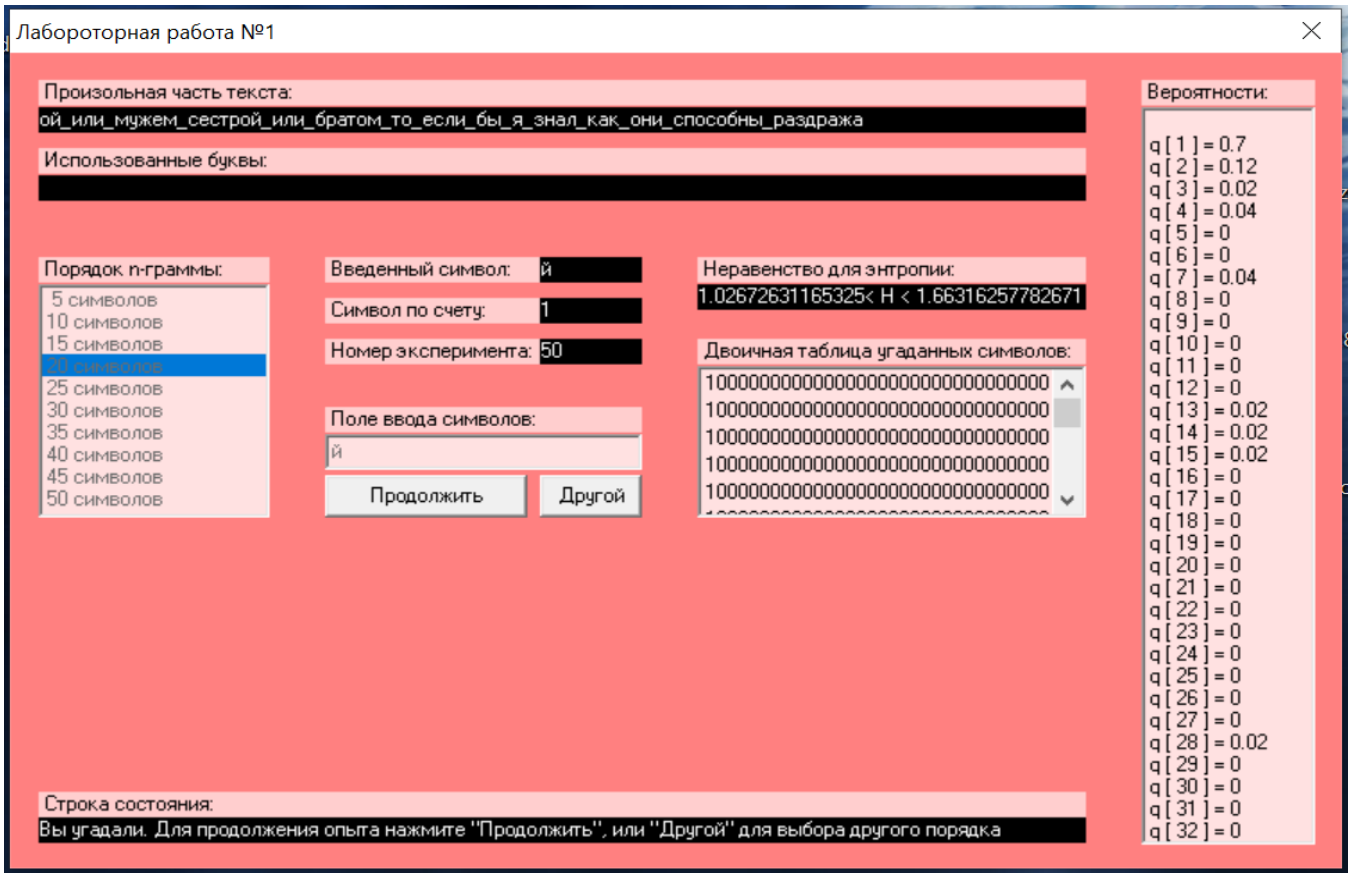
	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ы	ь	э	ю	я
	0	3716	17252	31125	10018	16840	8748	2987	10605	21133	38	21975	5670	14850	32798	27039	37753	11715	31905	17504	10293	1587	3566	777	12039	2503	337	0	0	7389	119	3987
а	39434	336	1748	5724	1547	3926	2906	2345	6450	290	2375	10232	22060	5451	6600	46	2499	5288	8564	9194	646	323	2259	396	2028	1177	746	0	0	28	2407	3966
б	623	2136	100	74	14	29	6992	40	5	5001	0	368	1974	39	706	5383	0	2764	304	28	2003	0	114	6	8	26	168	4674	298	59	27	971
в	13645	11776	17	204	43	467	10306	10	1028	6519	0	502	1112	254	3101	11359	435	1353	5130	746	1525	1	55	40	90	1160	13	5168	999	0	9	505
г	1726	3103	9	28	182	2984	1410	0	6	1511	0	251	3746	12	870	16099	0	2238	44	76	1327	10	5	0	81	14	0	2	0	24	0	0
д	2744	9392	65	2357	35	1724	11046	1290	35	6284	0	513	1257	137	3667	7323	296	2516	652	313	3876	2	104	679	26	110	0	1511	678	51	121	668
е	35839	219	2686	4292	6743	5080	3480	1725	2995	144	4057	2824	11709	7398	15125	299	1871	15819	7650	11758	252	156	1025	623	1843	1852	1206	0	0	3	188	506
ж	629	3262	22	9	40	1534	6110	29	0	3017	0	278	18	10	1779	547	0	16	14	7	504	0	0	0	172	0	0	0	84	1	0	0
з	3980	11097	260	1748	656	1663	888	262	186	1056	0	452	554	818	3022	1409	98	731	23	52	963	0	0	0	10	14	0	894	390	5	2	326
и	45738	309	1050	4385	1475	3326	3927	615	4785	1156	2162	4943	16249	5119	7397	634	768	1877	6749	7120	36	136	3336	1998	4125	1491	239	0	0	8	644	1841
й	15596	13	79	98	12	349	49	1	49	12	21	1322	47	194	487	135	44	73	1071	652	3	33	3	113	306	72	2	0	0	2	0	4
к	10617	15651	2	392	8	21	2011	26	17	7002	0	277	2223	8	1272	17617	4	4912	482	1543	4207	6	7	76	6	30	0	1	0	69	4	3
л	20263	14712	137	29	221	106	9444	590	128	17711	0	935	3461	92	835	13571	38	1	4564	290	3187	45	6	0	213	14	11	2282	8953	26	1854	3639
м	16165	8062	137	35	270	14	7141	0	0	6110	0	252	498	195	2886	8083	327	152	288	178	4338	20	9	21	141	1	2	2065	230	150	37	837
н	9876	22464	189	25	450	769	18262	76	238	15125	0	964	258	8	5021	21028	14	1166	1121	1452	6966	52	47	644	489	19	157	6604	2566	12	433	3082
о	44492	45	6918	13367	10323	11180	3533	4239	3053	1282	8143	4839	13398	13261	16712	884	2754	12375	14108	17571	556	211	1169	196	3806	2124	571	0	0	109	267	1721
п	231	4386	113	0	7	3	4365	0	0	2179	0	306	2024	1	538	22407	179	12053	159	456	1980	1	0	5	49	32	0	826	115	17	18	772
р	2455	15040	102	1056	268	1141	12316	742	84	13436	0	997	1328	678	2511	15459	150	824	744	1759	6382	24	386	167	202	399	68	4137	1297	293	246	1862
с	6588	2810	163	2048	91	598	6702	156	21	4022	0	6402	6942	2610	2141	5304	3761	452	1878	21965	1351	50	415	86	509	132	5	905	7844	64	312	9871
т	12196	11910	178	4599	26	255	10349	0	9	8012	0	1756	480	198	1988	31824	242	6888	2728	388	2890	24	36	226	313	38	10	4019	11501	617	118	1252
у	13805	230	2190	2177	2243	3294	354	2543	754	277	284	2429	5181	3083	994	61	1385	1138	3138	3518	185	53	994	5	2028	1107	687	0	0	269	2503	150
ф	103	325	5	0	0	0	280	0	0	533	0	10	94	8	5	743	0	265	29	27	279	44	0	0	22	1	0	29	17	14	5	0
х	6757	1183	3	653	11	0	169	0	0	413	0	13	390	92	599	4390	0	326	86	64	255	0	13	1	0	43	0	0	16	337	0	0
ц	668	1551	0	146	0	0	1747	0	0	923	0	13	0	0	3	590	2	0	0	46	368	0	0	0	0	0	0	511	0	0	0	0
ч	258	4339	0	8	0	0	6483	0	0	5489	0	1012	49	7	1985	184	0	10	0	7615	1644	1	1	0	77	265	0	4	476	4	0	0
ш	196	2021	2	108	0	0	3798	0	0	2832	0	1408	1447	88	476	746	30	53	0	294	493	0	0	21	1	25	0	0	1688	7	12	0
щ	13	880	0	0	0	0	2716	0	0	1964	0	0	0	0	88	5	0	3	0	0	533	0	0	0	0	0	0	0	43	0	0	0
ы	10103	0	670	2105	697	308	2670	112	209	33	3420	365	2394	2737	261	2	419	774	1348	1322	8	0	1801	16	278	1503	14	40	0	0	0	63
ь	23603	1	175	151	74	202	929	0	304	102	0	2127	0	366	2818	35	5	0	1885	198	0	38	0	369	497	1484	8	0	0	0	1047	775
э	11	0	17	11	52	1686	0	1	7	0	134	210	197	91	1314	0	22	371	72	5296	0	21	34	0	0	13	0	0	0	116	0	1
ю	5552	16	545	19	27	636	4	46	64	22	101	102	106	85	127	0	21	99	368	776	0	2	97	16	300	39	1397	0	0	0	120	0
я	22362	5	100	600	226	1335	232	247	526	48	108	417	2492	754	1451	0	107	301	1090	2862	9	0	334	87	253	60	604	0	0	0	193	217

H10



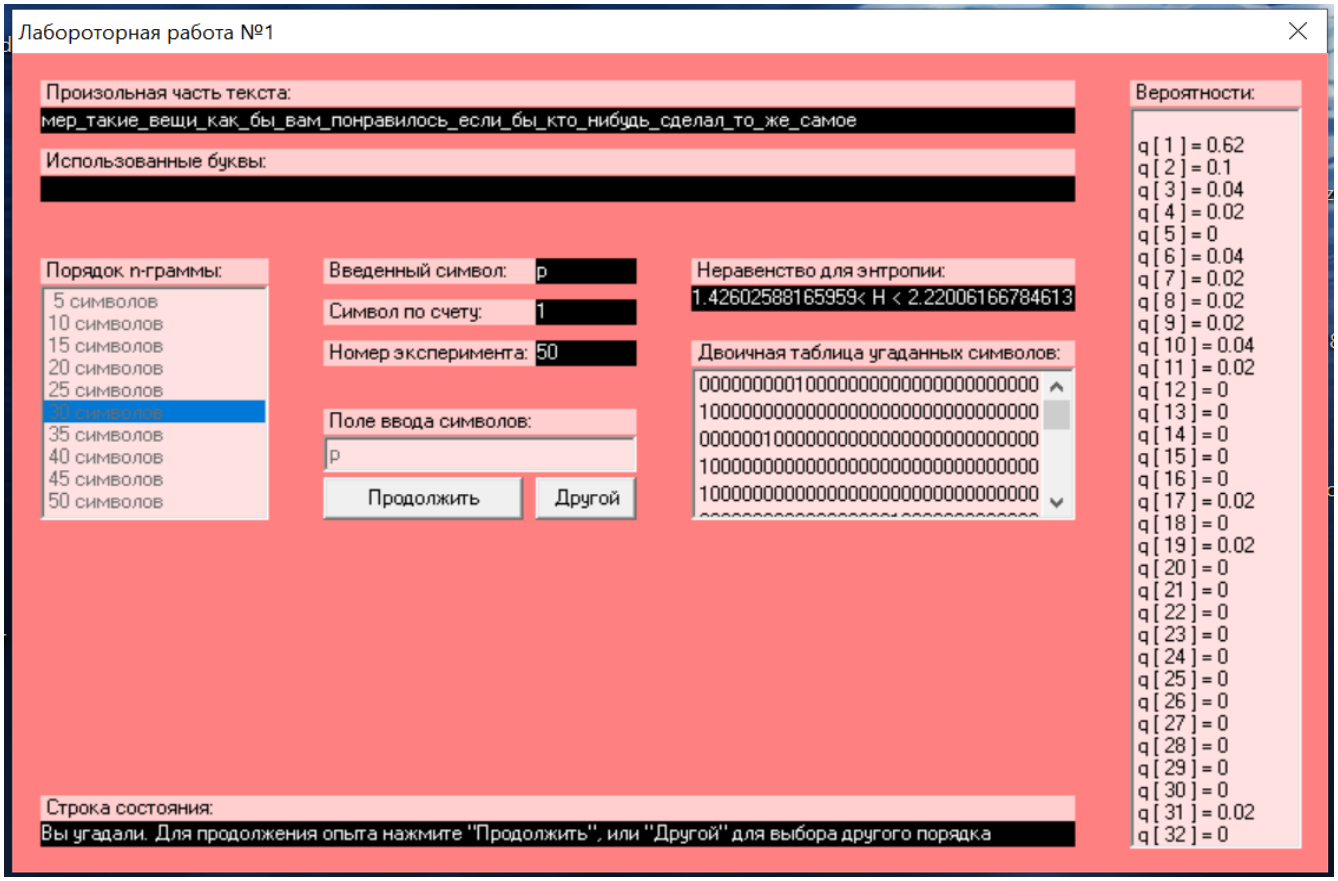
$$0.494688480055972 < R < 0.7176708727575805$$

H20



$$0.23756209939443898 < R < 0.3848195808139691$$

Н30



0.29061701170509674 < R < 0.4524375721425297

Ентропія на символ стаціонарного джерела визначається як

$$H_\infty = \lim_{n\to\infty} H_n$$

$$H_n = \frac{1}{n} H(x_1, x_2, ..., x_n)$$

$$H(x_1, x_2, ..., x_n) = -\sum_{z_1, z_2, ..., z_n} P(x_1 = z_1, ..., x_n = z_n) \cdot \log_2 P(x_1 = z_1, ..., x_n = z_n)$$

Надлишковість джерела відкритого тексту:

$$R = 1 - \frac{H_\infty}{H_0}$$

$$H_0 = \log_2 m$$

	Ентропія	**Надлишковість**
Без пробілів з шагом 1	4.167734	0.158747
Без пробілів з шагом 2	3.167661	0.360611
З пробілами з шагом 1	3.994742	0.201052
З пробілами з шагом 2	3994742	0.201052

Висновок: виконуючи цю лабораторну роботу, ми навчилися підраховувати частоти біграм та монограм, визначати ентропію на символ джерела та надлишковості джерела відкритого тексту.